

Glossary

Achievement Target. An achievement target is an educational goal to be attained through instruction, hence the prime focus of any assessment. The assessment is designed to measure for the achievement target: what evidence is there of the target being hit? Any performance task must therefore be designed “backwards” from the target to be valid and reliable. The following two questions must yield the same answer, in other words: “What does the task measure? What are the intended results of schooling or program related to this task?”

The targets are cast in general yet measurable terms (e.g. “students should be able to solve mathematical problems of multiple variables” or “students should have effective control over all key modes of communication”). Targets are thus different from what were once called behavioral objectives because targets are more comprehensive and complex. (The best assessment tasks, however, are those which are built upon a clear conception of discrete performance standards.)

Targets are generally of three kinds: understandings (which relate to the sought-for sophistication and thoroughness of student knowledge), competencies (which refer to intellectual skills and know-how), and mature habits of mind and attitudes (such as craftsmanship, perseverance, and tolerance of ambiguity). Educators often use the shorthand term “outcome” as a synonym for achievement target (see outcome).

Analytic trait. Analytic-trait (or primary-trait) scoring of performance involves the use of separate criteria in scoring work, typically involving separate rubrics for each key criterion. In effect, a performance is assessed numerous times, using the “lens” of a separate criterion each time. Analytic-trait scoring is thus in contrast with holistic scoring whereby a judge forms an overall impression about a performance.

Anchors. Anchors are the samples of work or performance used to set the specific performance standard for each level of a rubric. Thus, attached to the paragraph describing a 6-level performance in writing would be 2-3 samples of writing that illustrate what a 6-level performance is. (The anchor for the top score is often called the “exemplar.”)

It must be underscored that anchors are essential to reliability. A rubric without such anchors is typically far too ambiguous to set a clear standard. What, after all, do such phrases as “clear and persuasive” or “elegant mathematical solution” mean unless we have examples of work that give concrete and stable definition to them? Without anchors (or models) the assessment is incomplete at best and fatally flawed at worst: there will be a lack of inter-rater reliability since judges will be left to interpret the rubrics in a vacuum, students will not be able to self-assess or understand the expectations, and local judges will not be able to justify their standards as being credible.

The counter-intuitive fact is that the rubrics can only be fully developed after we have the anchors in hand. Otherwise, how would we be able to come up with valid and richly descriptive rubrics, unless we had already sorted the performances into levels or piles?

Anthology. An anthology is a representative and revealing collection of one’s work. Student anthologies serve two distinct purposes: providing a documentation of the student’s work, and serving as the basis for evaluation of work over time. The documentation typically serves three functions: revealing the student’s control over all the major areas/techniques/genres/ topics of the program, allowing students to reflect on and show off their best work, and providing evidence of how proficiencies evolved and were refined.

For evaluation purposes, anthologies can do what traditional testing cannot do: provide direct evidence for evaluating the student’s ability to make progress, over time, at mastering essential concepts, methods, and habits of mind.

In evaluating anthologies, judges must insure adequate reliability in scoring. A common procedure is to have meetings where people bring samples of the best, middle-level, and worst student work at which time grades are re-calibrated based on a consensus about standards. (The British and Australians refer to such re-calibration as “moderation” and it is a central feature of their assessment, the Advanced Placement scoring, and the New Standards exam system recently developed in this country.)

Anthologies differ from student portfolios in that in an anthology, clear and standardized criteria are used to determine what kinds of evidence must be available for evaluation. In many portfolio programs, students have great leeway in what does or does not go in or stay in.

Assess, assessment. To assess is to thoroughly and methodically analyze student accomplishment against specific goals and criteria. (The word comes from the Latin - assidere - meaning “to sit beside”). Assessment techniques include tests, exhibits, interviews, surveys, observation, etc. Good assessment requires a balance of techniques because each technique is limited and prone to error.

How does a “test” differ from an “assessment”? A test is one type of assessment -- typically the only mode of assessment for formal assessment in schools. The distinction is also partly one of manner and attitude, as implied by the Latin origin of the word: to assess is to ‘sit with’ the student. The implication is that in an overall assessment we should look beyond any individual test result to make a judgment about the student’s overall or habitual performance (making the judgment more reliable and valid). Put differently, in an “assessment” we report out the various facets of performance against criteria, and try to give useful feedback; in a test, we believe we can generalize from the single event adequately; in an evaluation, we place a value on that performance, not just a description.

Authentic assessment. An authentic assessment is composed of tasks and activities designed to simulate or replicate important, real-world challenges. The heart of authentic assessment is performance-based testing -- asking the student to use knowledge in a realistic way. Thus, the context of the assessment, not just the tasks (“messiness” of problem, ability to seek feedback and revise, access to apt resources, etc.), must be more realistic than conventional secure testing. Authentic assessments are meant to do more than “test”: they should teach students (and teachers) what the “doing” of a subject looks like and what kinds of performance challenges are actually considered most important in a field or profession. The tasks are chosen because they are representative of essential questions or challenges facing practitioners in the field.

An ‘authentic’ test directly measures students on the performances we value. By contrast, multiple-choice tests are indirect measures of performance. (Compare, for example, the road test versus the written test for getting a driver’s license. In the field of measurement, “authentic” tests are called “direct” tests.)

Note, however, that a task can be authentic but not be properly designed to allow for valid inferences. Simply because a task is hands-on, doesn’t mean that the task permits valid inferences for the goals in question. Similarly, a test could be “inauthentic” but valid, in that the inferences from the test do yield correlations with the direct test of performance. Thus, if the paper-and-pencil test for a driver’s

license correlates with driving records and accident rates, the paper and pencil test is valid even though it is an inauthentic test of driving performance.

Benchmark. A benchmark in an assessment system is often defined as a developmentally-apt content or performance standard, sometimes called a “milepost” standard. In many district-wide systems, there are “benchmarks” set for Grades 4, 8, 10, and 12, for example.

But in industry, “benchmark” is often used as a verb, defined as the search for a best performance or achievement specification for a particular objective found anywhere in the world. The resulting “benchmark” (noun) sets the highest possible standard of performance, a goal to be met or exceeded locally. Thus, a benchmark in this sense is used when we want our assessment to be anchored by the best possible samples of work (versus anchored by samples of work from an average school district.)

An assessment anchored by benchmarks, in either sense of the word, should not be expected to yield a predictable “curve” of results (true for any true criterion-referenced test). Standards differ from reasonable expectations. (See standard.) We might get very few products or performances -- or even none at all -- that match the benchmark performance.

Bias. A judge of student performance is technically biased if a) his/her standards are higher or lower than the agreed-upon ones, b) he/she tends to focus on particular strengths or weaknesses in the performance that are not consistent with specified guidelines, or c) he/she focuses on the performer’s personal traits or characteristics vs. the qualities of the performance or product to be judged.

Thus, a judge who consistently gives B's to papers that would ‘normally’ get an A from trained readers is ‘biased’. Or a judge who consistently knocks off more points for spelling errors than the scoring system suggests is ‘biased’. (Note, though, that if the bias is consistent, the judge is reliable.)

A judge who is biased against certain styles or kinds of performances -- or, worse certain kinds of performers -- is ‘biased’ in the more colloquial sense. Both forms of bias creep into judgment-based scoring of work and need to be guarded against. The problem of bias (and unreliability) leads many people to favor multiple-choice testing since the scoring on such tests is inherently free from bias. This overlooks, however, the fact that the design of test items is open to bias. The

proposed assessment system builds in a variety of oversight features to minimize bias.

Bloom's Taxonomy. Over forty years ago, Benjamin Bloom and his colleagues developed a schema for distinguishing the simplest forms of recall from the most sophisticated uses of knowledge in designing student assessments. The six elements were called Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation.

To speak of 'higher-order' thinking is to refer to the last three stages of the taxonomy. Note, then, that "application" is not a higher-order skill. This seems contradictory, since many advocates of authentic assessment talk about getting the student to more effectively "apply" knowledge. But this is not what Bloom and his colleagues meant by "apply." They were speaking of those cases where one kind of knowledge or skill must be used, as in a typical vocabulary test where a sentence must be constructed or a math word problem -- not of the more sophisticated act of drawing upon a repertoire to solve a complex, multi-faceted problem. The authors' description of "synthesis" better fits the meaning of "performance testing" since they stress that such an aim requires the "students' unique production."

Criteria. Criteria are the qualities that must be met by performances/ products for work to be up to standard and for the performances to be deemed successful. To ask "what are the criteria?" amounts to asking: "Where should we look in examining students' products or performances to know if they were successful? How well they did? What should we look for? For what kinds of errors will we then take points off, and to what degree?"

Criteria logically precede the design of specific performance tasks. If we are going to design a task that measures for "critical thinking," we need to know beforehand what the signs of such thinking are, and the traits of performance we need the task to measure for, hence build into the specific task.

The assessment must also determine how much weight each criterion should receive relative to other criteria in making our judgment. Thus, even if we agree that spelling and development of ideas are both important in judging writing, we must ask: what percentage should we assign to each?

The criteria used in judging performance, like a test itself, can thus be valid or invalid, and authentic or inauthentic. For example, one can assign students to do some original historical research (an authentic task), but grade the work only on

whether 4 sources were used and whether the report is exactly 5 pages long. Such criteria would be invalid because we can easily imagine a piece of historical work not meeting those two criteria but still being excellent research. Criteria should correspond to the qualities of masterful performance.

The most common error in the design of performance tests is to use criteria which over-stress mere form, content, and process (Is the paper organized? Are the facts accurate? Was the writing process used?) instead of highlighting criteria related to impact and thus to audience/purpose (Was the paper engaging? Memorable? Insightful?) Or the assessment mistakenly turns qualitative criteria (the aptness and thoroughness of the research) into quantitative criteria (there must be 3 different sources used).

In CLASS work, four different types of criteria are referred to: impact, work quality and craftsmanship, methods and manner, content validity, and sophistication. Impact criteria relate to effectiveness (e.g. persuasive, satisfying), work-quality-related to the care, polish, look, or order (e.g. organized, attractive, well-designed), method-related criteria relate to the quality of the process (e.g. efficient, methodical, clever), content criteria relate to the aptness and accuracy of the material or knowledge (e.g. accurate, apt, focused), and sophistication refers to the level of expertise or complexity used.

Impact is clearly at the heart of what we mean by performance, i.e. did the performance work? What was its effect, its result -- its outcome, irrespective of effort, attitude, and approach.

Methods and manner refers to the means, processes, attitude, or approaches taken in performance or in preparation for performance. Was the work efficient? Was the manner fluid and poised? Was the performer dedicated? These are method-related questions.

Work-quality refers to the attention to detail, polish, and organization taken in performance and rehearsal. Was the paper organized? Was the speech well-delivered? Was the lab write-up error-free? and done according to proper format? -- these questions are all about work quality

Content validity refers to the aptness, adequacy, or accuracy of the material or content.

Sophistication refers to the level of expertise displayed. Note that the content could be accurate but unsophisticated and vice versa -- hence the distinction between “valid content” and “sophistication.”

Direct and indirect tests. Direct tests measure the performer’s ability on the target or desired performance; indirect tests use (often deliberately simplified) ways of measuring the “same” performance out of context. Thus, a multiple-choice test of any complex performance (reading, writing, problem-solving) is, by definition, indirect. The ACT’s and SAT’s are “indirect” ways of assessing likely success in college. A “direct” test is therefore more authentic than an indirect test, by definition. However, an indirect test of performance can be valid: if results on the indirect test correlate with results on direct tests, then the test is valid by definition.

Genres of performance. A genre is a type or category of intellectual performance or product. For example, we speak of genres of writing (narrative, essay, letter) or speaking (seminar discussion, formal speech, giving directions). A genre is thus a sub-set of the three main modes of intellectual performance: oral, written, displayed.

Holistic scoring. A holistic score is the result of an overall impression of the quality of a performance. Holistic scoring is distinguished from analytic trait scoring, where separate rubrics are used for each separate criterion that makes up an aspect of performance.

There can be multiple holistic scores, however. There could be scores for an oral performance and a written performance, without breaking down those scores into the analytic components of each mode (e.g. the persuasiveness and clarity of the oral performance.)

Ill-structured. A question or task is ill-structured if there is no recipe or obvious formula to answer it; or if the task does not state or imply a specific strategy guaranteed to yield success. Such questions thus demand more than knowledge: they demand good judgment and imagination. All good essay questions, science problems, or design challenges are thus ill-structured: even when you understand the goal or know what is expected, you have to scratch your head and build not

merely an answer but a procedure as you go. Invariably, ill-structured tasks require constant self-assessment and revision, not just a simple application of knowledge.

Most ‘authentic’ problems are ill-structured; most test ‘items’ are not. Test questions are ‘well-structured’ in that there is a single, unambiguous right answer -- or an obvious procedure for solving it. Such ‘items’ are fine for validly assessing elements of knowledge but not appropriate for judging the student’s ability to use knowledge wisely -- namely, how to judge which knowledge and skill to use when. (Think, for example, of the differences between the ‘test’ of each drill in basketball, and the ‘test’ of playing the game well in performance: the drill is predictable and structured; the game is unpredictable and not scriptable.)

Longitudinal assessment. Longitudinal assessment (or “developmental” assessment) involves assessing the same performances over numerous times, using a fixed scoring continuum, to track progress (or lack of it) toward a standard. For example, the National Assessment of Educational Progress (NAEP) uses a fixed scale for measuring gains in mathematics performance over the 4th, 8th, and 12th grade. Similarly, the American Council on the Teaching of Foreign Languages (ACTFL) uses a novice-expert continuum for charting the progress of all language students over time. Almost all school testing, whether done locally or state-wide, is not longitudinal since the tests are one-time events with one-time scoring systems. The proposed assessment system will use scoring scales and tasks that can be used across many grade levels so as to provide longitudinal assessment.

Matrix sampling. Matrix sampling is a strategy used in large-scale testing programs to get better information about overall performance by using a variety of tests during the same test administration. For example, in a writing test, we might identify seven different genres of writing in which we wish students to develop some proficiency. In the test, each student might get only one writing task focused on one genre. But across all students tested, all genres would be tested. Matrix sampling thus gives us far richer and more valid information about how well program objectives are being met than if we gave the same test to all students.

While matrix sampling enables us to draw valid conclusions across schools, districts, and states for all the genres, it of course does not allow us to make generalizations for each and every student about their control of all genres. In other words, student scores are not universally comparable: a student score on one genre can only be compared against other student performances in the same genre. (This

is because success in one genre does not correlate with results in another genre: excellent poets do not necessarily write excellent analytic essays.)

“Stratified matrix sampling” means that if we test only a sample of students across all genres/topics/skill/knowledge, then the sample of students must accurately represent the full range of the student population based on whatever trait we want to highlight in our sample: Grade Point Average, prior tests scores, socio-economic status, etc.

Mode, performance mode. Mode refers to the type or method of performance being demonstrated and assessed. Modes of performance include: written, oral, visual. Within each mode there are many genres. (See genres).

Outcome. An outcome in education is shorthand for “intended outcomes of instruction.” An “intended outcome” is a desired result, a specific goal to which we commit as educators. In this book, we use the term “achievement target” to describe such intents.

To operationalize outcomes we have to agree on the specific “standards and measures” -- the tasks, criteria and standards by which we can assess whether what we intended has occurred. The word “outcome” is neutral. It does not imply a specific set of goals or values about education. It is used to describe the particular measurable goals that will be targeted in a particular system. So-called Outcomes-Based Education (OBE) is different: a specific approach to school reform that requires faculties to agree to agree about educational outcomes, which prescribes certain strategies for achieving consensus about outcomes, and which encompasses non-academic as well as academic objectives.

Open-ended tasks or questions. A task is open-ended if it does not lead to a single “right” answer. This does not imply that all answers are of equal value, however. Rather, it implies that many different acceptable answers are possible. Such answers are thus “justified” or “plausible” or “well-defended” as opposed to “correct.” Essay test questions, for example, are all open-ended. By contrast, all multiple-choice tests are not open-ended by design.

Perform, performance. To perform is to “act upon and bring to completion.” To perform in the intellectual realm involves using one’s knowledge to effectively act or bring to fruition a complex product in which one’s knowledge and expertise is

revealed. Music recitals and auto mechanic competitions are performances in both senses; so are oral exams.

Many educators mistakenly use the phrase “performance assessment” when they really mean “performance test” (see assess, assessment, above). In a performance assessment we would use more than a single test of performance and we might well use other modes of assessment as well (such as surveys, interviews of the performer, quizzes, etc.)

The use of the word “performance” highlights an important difference between authentic and multiple-choice tests. An authentic test of performance is more than the sum of isolated drills in that area of performance. Most conventional short-answer or multiple-choice tests are more like the drills in sports than the ‘test’ of performance. Real performers -- be they athletes, debaters, dancers, scientists or actors -- must learn to ‘perform’ with knowledge, to use it wisely and effectively in complex settings. They must use their judgment as well as their knowledge. By contrast, multiple-choice tests merely ask the student to recall, recognize or “plug in” isolated ‘items’ of finished knowledge or skill, one at a time. Since many types of performance are ephemeral actions, a fair and technically sound assessment typically involves the creation of products. This insures adequate documentation and the possibility of appropriate review and oversight in scoring the performance.

Portfolio. A portfolio is a representative collection of one’s work. As the word’s roots suggest (and as is still the case in the arts), the sample of work is fashioned for a particular objective and carried from place to place for inspection or exhibition.

In academic subject areas such as English/Language Arts or Mathematics, a portfolio often serves two distinct purposes: providing a documentation of the student’s work, and serving as the basis for evaluation of work-in-progress or work over time. The documentation typically serves three functions: revealing the student’s control over all the major areas/techniques/genres/topics of the course or program, allowing students to reflect on and show off their best work (by letting them select which works will be put in the portfolio), providing evidence of how works evolved and were refined. (See *Anthology*.)

Process. In the context of assessment, process refers to the intermediate steps the student takes in reaching the final performance or end-product specified by the

assessment. “Process” thus includes all strategies, decisions, sub-skills, rough drafts and rehearsals used in completing the given task.

In being asked to evaluate the ‘process’ leading to the final performance or product, the assessor is sometimes asked to explicitly judge the student's intermediate steps, independent of what can be inferred about those processes from the end result. For example, one might be asked to separately rate a student’s ability to work with a group or in doing pre-writing as part of a research project -- independently of the ultimate product the group or individual writer produces. We should beware of routinely scoring ‘process’ separately, however, even if in teaching we want to assess those ‘process’ skills. After all, as the word “performance” implies, the emphasis is on whether the final product or performance met the standards set -- irrespective of how the student got there.

Product. A product is the tangible and stable residue of a performance and the processes that led to it. The product is valid for assessing the student’s knowledge to the extent that success or failure in producing the product a) is dependent upon the knowledge we taught and want to assess, and b) appropriately ‘samples’ from the whole curriculum in a way that mirrors the relative importance of the material in the course (as reflected in the test blueprint).

Reliable, reliability. Reliability in measurement and testing refers to the accuracy of the score. Is it sufficiently free of error? What is the likelihood that the score or grade would be constant if the test were re-taken or the same performance were re-scored by someone else? Error is unavoidable; all tests, including the best multiple-choice tests, lack 100% reliability. The aim is to minimize it to tolerable levels.

Score error is not a defect in the test-maker’s methods or double-checking, but a statistical fact about how extraneous factors inevitably influence test-takers or judges; or the limits of using a small sample of questions or tasks in a single sitting. Instructions are mis-read; the wrong box is marked; judges disagree, sometimes for good reasons; one day’s performance turns out to be an exception to a rule. That is why we say on the “objective” SAT’s, for example, that the student's “true” score is within a range -- in this case, plus or minus 35 points from the score received.

In performance tests the reliability problem typically occurs in two forms: 1) to what extent can we generalize from the single or small amount of performances to the student’s likely performance in general? Is the score truly representative of

the student's general capacities and patterns of results? and 2) what is the likelihood that different judges will see the same performance in the same way? The second question involves what is typically termed inter-rater reliability.

How can we obtain adequate reliability of both kinds? First, we ensure that there are multiple tasks for the same outcome: better reliability is obtained when we have many tasks and multiple judges. Also, with well-trained and supervised judges, working from specific anchor papers/performances and with clear rubrics greatly improves reliability. (These procedures have long been used in the Advanced Placement Program.)

Rubric. A rubric is a scoring guide which enables judges to make reliable judgments about student work and students to self-assess. The rubric answers the question: What does mastery (and varying degrees of mastery) for an achievement target look like?

A typical rubric:

1. is based on a continuum of performance quality, built upon a scale of different possible score points to be assigned. Scores often range from 6 as the top score, down to 1 or 0 for the lowest score;
2. identifies the key traits or dimensions to be examined and assessed (e.g. "syntax" or "understanding of scientific method"); and
3. provides key features of performance for each level of scoring, in "descriptors," which signify whether or not the criteria have been met, and to what extent -- thus enabling the judge to assign the right score.
4. provides "indicators" under or within each descriptor to provide concrete signs or suggestions for more accurate scoring and self-assessment.

Sample, sampling. All testing involves the act of sampling from a vast domain of possible knowledge and tasks. Like the Gallup polls, sampling enables the assessor to draw valid inferences from a limited inquiry -- if the sample is apt and justified.

Two different kinds of “sampling” go on in test design: sampling with respect to the wider domain of all possible curricular questions, topics and tasks; and testing only a sample (or sub-set) of the entire student population instead of testing everyone. These two kinds of sampling get combined to form matrix sampling whereby we test many or all students using different versions of the test to cover as much of the domain of knowledge as possible.

When we try to sample the domain of a subject through a small task, we must ask: what feasible and efficient sample of tasks or questions will enable us to make valid inferences about the student’s overall performance (since we cannot possibly test the student on everything that was taught and learned)? When we try to use a sub-set of the population to construct a more efficient and cost-effective approach to testing we are asking the question the pollsters ask: what must be the composition of any small sample of students so that we can validly infer conclusions about the system-wide performance of all students using the results from our small sample?

Scale, scoring scale. The scoring scale is the equally-divided up continuum (number-line) for scoring performance. The scale identifies how many different scores, from high to low number, which will be used. Performance assessment typically uses a much smaller scale for scoring than standardized tests. Rather than a scale of 100 or more, most performance-based assessment uses a 6-point scale; rarely does a scale contain more than 10 points.

There are two inter-related reasons for this use of a small number of score points. Each place on the scale is not arbitrary (as it is in norm-referenced scoring); it is meant to correspond to a specific criterion or quality of work. The second reason is practical: to use a scale of so many discrete numbers makes reliability unlikely, and attempts at such fine criterion-referenced distinctions become picky or arbitrary.

Scoring Guide. See *Rubric*.

Secure test, security. A test is secure when teachers and/or students do not have prior access to the test for purposes of preparation. Most multiple-choice tests must be secure or their validity is compromised since they rely on a small number of questions with ‘correct’ answers provided within the stated four or five choices. Interestingly, many valid performance assessments are not secure. The student to be assessed knows the musical piece, debate topic, oral exam questions, or term

paper subject in advance; the teacher/coach fairly ‘teaches to the (known) test’ of performance.

Standard. To ask: “What is the standard on this assessment?” is to wonder about how well the student must perform to do well or adequately. A performance standard is a specific result or level of achievement that is deemed exemplary or appropriate.

But confusions abound: the word is sometimes used in education as a synonym for “high expectations.” At other times, “standard” is used as a synonym for “benchmark” -- the best the performance or product can be done or has been done. And in large-scale testing, “standard” has often implicitly meant “minimal standard.” One can also often hear standards discussed as if they were general guidelines or principles. Often, speakers confuse “content” standards with “performance” standards. Finally, “standards” are routinely confused with the “criteria” for judging performance: many people falsely believe that a rubric is sufficient for scoring performance.

In the work of CLASS, we typically make reference to four types of standards: content, competency, performance, and work design:

- a. Content Standards: What students should know
- b. Competency Standards: What students should be able to do
- c. Performance Standards: How well students must do their work
- d. Work Design Standards: What worthy work students should encounter and do well

1) A “standard” is not necessarily an “apt high expectation.” A standard is a standard -- whether or not most people can or cannot meet it. That is very different than where we believe a good number of students not only can but ought to meet it, if they persist and get good teaching from teachers with high expectations. When we set “high expectations” we typically mean in practice that a student should get a 4 or better on a 6-point performance assessment scale, for example.

2) A standard is set by an “exemplary” anchor performance, but not necessarily by a national “benchmark” for the same reasons as in the previous point. Consider wider-world benchmarks: the 4-minute mile. The Malcolm Baldrige Award winning companies. Hemingway. Peter Jennings. Few student performers, if any, will meet such “benchmarks” standards. Standards in this sense are worthy targets -- but perhaps not realistic objectives. In school tests we rarely set performance standards using such benchmarks. The standard is typically set through the selection of developmentally-appropriate or experience-level-appropriate anchors or exemplars of performance. The choice of such exemplary work samples sets the de facto standard.

The key assessment question then becomes: where should the samples of student work come from? What would be a valid choice of anchors? What we typically do is select the best work available from the overall student population being tested. (We believe, however, that the students need to be more routinely tested with anchors that come from slightly more advanced/experienced students, to serve as a helpful longer-range target, and to provide on-going feedback.)

3) A “standard” differs from the “criteria” by which we judge performance. The criteria for the high jump or the persuasive essay are more or less fixed -- no matter the age or ability of the student. All high jumps, to be successful, must meet the same criterion: the bar must stay on. In writing, all persuasive essays must use lots of appropriate evidence and argument effectively. But how high should the bar be? How sophisticated and rigorous should the arguments be? Those are “standards” questions. (A rubric typically refers in the descriptors to both criteria and standards.)

4) Performance standards have been traditionally operationalized through fixing a minimally-acceptable performance level through so-called “cut-off” or “cut” scores. Typically, in both classroom grading and on state tests, a 60 is considered a minimal “standard” of performance. But test designers are rarely asked to establish a defensible “cut” score. Stating at the outset that 60 is passing and 59 is failing is arbitrary: Few tests are designed so that there is a qualitative difference between a 59 and a 61. It is thus all too easy, when thinking of a ‘standard’ as a cut-off point, to turn what should be a criterion-referenced scoring system into a norm-referenced scoring system.

5) Improving “content” standards will not necessarily raise “performance” standards. Content refers to input and performance to output. Content standards state the particular knowledge the student should master. Many current reforms assume improving the “inputs’ will necessarily improve the output. But this is clearly false: one can still receive poor quality work from students in a demanding course of study. In fact, it is reasonable in the short term to obtain worse performance by raising content standards only: higher standards merely in the difficulty of what is taught will likely lead to higher failure by students, if all other factors (teaching and time spent on work) are held constant.

The key question to ask in setting valid and useful performance standards must always be: at what level of performance would the student be “appropriately qualified or certified”? An effective solution to operationalizing standards is thus to equate (internal) teacher and school standards to some equivalent, worthy level of achievement in the outside world -- a wider-world benchmark -- thus lending substance, stability, and credibility to our scoring. This is a common feature of vocational, musical, athletic and other performance-based forms of learning.

Standardized. A test or assessment is standardized if the administrative conditions and protocol are uniform for all students. In other words, if all students face similar logistical, time, material, and feedback guidelines and constraints, then the test is standardized.

There are three common misconceptions about standardized tests:

1. It is erroneous to say a multiple-choice test is synonymous with “standardized test.” A performance task, administered uniformly as a test, is also a standardized test -- as we see in the road test for a driver’s license or a qualifying meet for the Olympics.
2. It is a common but mistaken view to argue that “standardized” tests are always “objectively” (i.e. machine) scored. But the Advanced Placement Exam Essays and all state writing tests are scored by judges – yet standard in their administration.

3. Many argue as if only national norm-referenced or criterion-referenced tests (such as the SAT) can be “standardized” but clearly a departmental exam in a high school is also a standardized test.

An important implication, then, is that by definition, all formal tests are “standardized.” This is not true of an assessment, however. In an assessment, the administrator is free to vary the questions, the tasks, the order of the tasks, the time allotted, etc. (i.e. the protocol) in order to be satisfied that the results are fair, valid and reliable. This was the argument made by Piaget for his “clinical method” as opposed to the “test method” of Binet. See assessment.

Task, Performance Task. A task is a complex challenge, which requires a multi-faceted performance, implying a performance standard, and which requires work to be developed through different stages and perhaps across different “content.” (The British use the phrase “integrated task” to capture this idea). It typically involves diverse activities spread out over a lengthy period of time: doing research and bringing it to fruition in reports, building a museum exhibit, etc. A performance task thus demands that we bring to bear a repertoire of knowledge and skill to solve a problem or through a series of judgments and actions. Most tasks are goal-directed: they are ‘done’ when we have successfully fashioned a performance or product to specifications. A task thus differs from a conventional test item in the same way that “successfully building a balsa bridge to withstand y pounds per sq. inch” differs from solving physics textbook paper and pencil problems.

Valid, validity. Validity concerns the inferences from tests about a student’s ability and their aptness. All tests are brief samples of work: does the test design correspond to the ‘blueprint’ of the priorities entire course syllabus or program? Do the test results correlate with other performance results we have faith in? Does the small sample of questions accurately correlate with what students would do if we tested them on everything that was taught in the course? Do the results have predictive value, i.e. do they correlate with likely future success in the subject in question? Some or all of these questions must have a “yes” answer for the test to be valid.

To be precise, therefore, it is not a test itself that is valid but the inferences that we claim to be able to make from the test results. Thus, the purpose of the test must be considered when assessing validity. Multiple-choice reading tests may well

be valid if they are used to test the student's comprehension ability or monitor grade-level reading ability of a district's population as compared to other large populations. They may not be valid as measures of a pupil's repertoire of reading strategies, and the ability to construct apt and insightful responses to texts.

The format of the test can be misleading, in an important sense, therefore: an inauthentic test can still be technically valid. It may aptly sample from the subject domain and predict future performance accurately but nonetheless be based on inauthentic, even trivial tasks. The SAT college admissions test and tests such as the Otis-Lennon School-Ability Index are said by their makers to be valid in this more limited sense: efficient proxies that serve as useful predictors.

Conversely, an authentic task may not be valid. Can we accurately and reliably predict from the specific task performed that the student has control over the entire domain? Does one type of task enable us to infer to other types of tasks (say, one genre of writing to all others)? No. Thus, there can often be inadequate "generalizability" from the (typically few) tasks used in performance assessment. One solution is to use a wide variety of student work of a similar type or genre, collected over the year, as part of the summative assessment.

The scoring system can raise other questions about validity. To ask if a performance-based test is "valid" is to ask, within the limits of feasibility, if we are scoring the most important aspects of performance as opposed to the most easily-scored aspects. Have we identified the most apt criteria and the most apt difference in quality? Or have we merely scored what is easy to count and score? (Have we sacrificed validity for reliability, in other words?) Large-scale tests typically make arbitrary decisions about the kinds of criteria to be used in assessment and the weighting of each criterion in reference to the others (e.g. 25% each for four criteria, even though in reality the criteria are not viewed equally.)